

Perspectives on Deafness

*Series Editors*

Marc Marschark

Patricia Elizabeth Spencer

*The World of Deaf Infants: A Longitudinal Study*

Kathryn P. Meadow-Orlans, Patricia Elizabeth Spencer,  
and Lynn Sanford Koester

*Sign Language Interpreting and Interpreter Education:*

*Directions for Research and Practice*

Edited by Marc Marschark, Rico Peterson, and Elizabeth A. Winston

*Advances in the Spoken Language Development of Deaf and Hard-of-Hearing Children*

Edited by Patricia Elizabeth Spencer and Marc Marschark

*Advances in the Sign Language Development of Deaf Children*

Edited by Brenda Schick, Marc Marschark, and Patricia Elizabeth Spencer

ADVANCES IN THE

## Spoken Language Development

OF DEAF AND HARD-OF-HEARING CHILDREN

EDITED BY

Patricia Elizabeth Spencer and Marc Marschark

OXFORD  
UNIVERSITY PRESS

2006

**OXFORD**  
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further  
Oxford University's objective of excellence  
in research, scholarship, and education.

Oxford New York  
Auckland Cape Town Dar es Salaam Hong Kong Karachi  
Kuala Lumpur Madrid Melbourne Mexico City Nairobi  
New Delhi Shanghai Taipei Toronto

With offices in  
Argentina Austria Brazil Chile Czech Republic France Greece  
Guatemala Hungary Italy Japan Poland Portugal Singapore  
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Copyright © 2006 by Patricia Elizabeth Spencer and Marc Marschark

Published by Oxford University Press, Inc.  
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press

All rights reserved. No part of this publication may be reproduced,  
stored in a retrieval system, or transmitted, in any form or by any means,  
electronic, mechanical, photocopying, recording, or otherwise,  
without the prior permission of Oxford University Press.

Library of Congress Cataloging-in-Publication Data  
Advances in the spoken language development of deaf and hard-of-hearing children /  
edited by Patricia Elizabeth Spencer and Marc Marschark.

p. cm.—(Perspectives on deafness)  
Includes bibliographical references and index.

ISBN-13 978-0-19-517987-3

ISBN 0-19-517987-0

1. Deaf children—Language.
  2. Hearing impaired children—Language.
  3. Language awareness in children.
  4. Language acquisition.
  5. Oral communication.
- I. Spencer, Patricia Elizabeth. II. Marschark, Marc. III. Series.  
HV2391.A38 2005  
401'.93—dc22 2004027122

## A Computer-Animated Tutor for Language Learning: Research and Applications

*Dominic W. Massaro*

This volume documents how successfully humans learn and use language without adequate auditory input. Sign language parallels spoken language in acquisition, use, and communication, but even oral language can serve communication when the auditory input is degraded or even absent. Lipreading (speechreading because it involves more than just the lips) allows these individuals to perceive and understand oral language and even to speak (Bernstein, Demorest, & Tucker, 2000; Kisor, 1990; Mirrelles, 1947). Speechreading seldom disambiguates all of the spoken input, however, and other techniques have been used to allow a richer input. Cued speech, for example, is a recent deliberate solution to having a limited auditory input and consists of hand gestures while speaking that provide the perceiver with disambiguating information in addition to what is seen on the face. Other devices such as vibratory aids that transduce the auditory speech into tactile input have also been used (Bernstein, Demorest, Coulter, & O'Connell, 1991; Waldstein & Boothroyd, 1995). For hard-of-hearing individuals, however, processing oral language from the voice and face is a natural solution to handling two limited input channels. This situation is the focus of this chapter, and the technology, research, and pedagogy described here support the premise that all humans easily exploit multiple sensory inputs in language processing.

The need for language tutoring is pervasive in today's world. Millions of individuals have language and speech challenges, and these individuals require additional instruction in language learning. For example, it is well known that hard-of-hearing children have significant

deficits in both spoken and written vocabulary knowledge (Breslaw, Griffiths, Wood, & Howarth, 1981; Holt, Traxler, & Allen, 1997). A similar situation exists for autistic children, who lag behind their typically developing cohort in language acquisition (Tager-Flusberg, 2000). Currently, however, these needs are not being met. One problem that the people with these disabilities face is that there are not enough skilled teachers and professionals to give them the one-on-one attention that they need. So they resort to other resources, such as books or other media, but these are not easily personalized to the students' needs, lack the engaging capability of a teacher, are rather expensive, and are relatively ineffective.

In addition to these individuals with specific language challenges, many other persons must learn a new language. Given the highly mobile society, individuals of all walks of life find themselves in situations in which successful business and social interactions require use of a nonnative language. As an obvious example, English is becoming increasingly necessary and desirable worldwide, and the number of people in the world who are learning English is increasing at a rapid rate. One of our goals is to apply the knowledge that has been obtained in speech science and related disciplines to several domains of language learning. These include the learning of vocabulary and grammar as well as the perception and production of speech.

This goal was facilitated or even created by our serendipitous relationship with a computer-animated talking head, Baldi<sup>TM</sup> (Dominic W. Massaro, Santa Cruz, CA, USA). We first incorporated Baldi in order to control the visible speech presented to participants in our research on multisensory or multimodal speech perception. Baldi's versatility soon convinced us that he had potential value as an embodied conversational agent who could guide students through a variety of exercises designed to teach vocabulary and grammar and to develop linguistic and phonological awareness. We believe this vocation for Baldi holds promise for all children with language challenges and even for typically developing children. Baldi is now used in the Language Wizard and Tutor program (described in Bosseler & Massaro, 2003), which encompasses and implements developments in the pedagogy of how language is learned, remembered, and used.

Given this context, this chapter continues with the evidence for the value of multimodal linguistic input followed by a theoretical description of multimodal speech perception. The technology behind Baldi's attractive exterior is described, as well as his development and the evaluation of his effectiveness in simulating a naturally talking person. The case is then made for the importance of vocabulary in cognitive development, the value of the direct teaching of vocabulary and grammar, and the development and evaluation of the Language Wizard/Tutor for language tutoring.

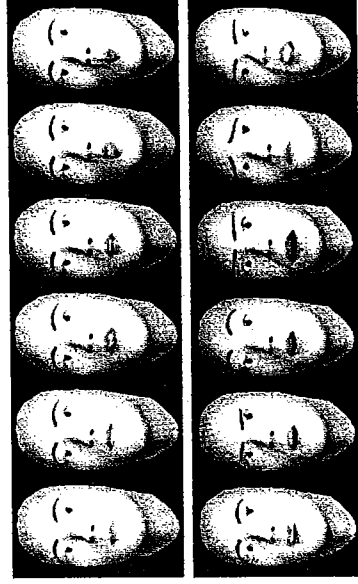


Figure 10-1. Baldi, our three-dimensional computer-animated talking head, has realistic speech as well as convincing emotions. The top panel shows six phoneme articulations, and the bottom panel, six basic emotions. In addition, Baldi has a sister, Baldette, who has both hair and a body with appropriate gestures of oral language.

## MULTIMODAL SPEECH AND ITS VALUE FOR LANGUAGE TUTORING

Speech science evolved as the study of a unimodal auditory channel of communication because speech was viewed as solely an auditory event (e.g., Denes & Pinson, 1963). There is no doubt that the voice alone is usually adequate for understanding and, given the popularity of mobile phones, might be the most frequent medium for today's communication. However, speech should be viewed as a multimodal phenomenon because the human face presents visual information during speech production that is critically important for effective communication. Experiments indicate that our perception and understanding are influenced by a speaker's face and accompanying gestures, as well as the actual sound of the speech (Massaro, 1987, 1998, 2000, 2004; Summerfield, 1987).

In face-to-face communication, visible speech from the talker's face (or from a reasonably accurate synthetic talking head) improves understanding. For many individuals with severe or profound hearing loss, understanding visible speech is essential to orally communicating effectively with others. Even for typically hearing individuals, the face is valuable because many communication environments involve a noisy auditory channel, which degrades speech perception and recognition. Visible speech is also an important oral communication channel for individuals with specific limitations in processing auditory information, or with other types of language challenges. One of the central themes of our research is that viewing speech as a multimodal experience can also improve language tutoring, an important challenge for applications of speech science.

## BALDI AND THE VALUE OF VISIBLE SPEECH

The value of visible speech in face-to-face communication was the primary motivation for the development of Baldi, a three-dimensional computer-animated talking head, shown in figure 10.1. Baldi provides realistic visible speech that is almost as accurate as a natural speaker (Cohen, Beskow, & Massaro, 1998; Massaro, 1998, ch. 13). Baldi's visible speech can be appropriately aligned with either synthesized or natural auditory speech. Baldi also has teeth, a tongue, and a palate to simulate the inside of the mouth, and the tongue movements have been trained to mimic natural tongue movements (Cohen et al., 1998). We have also witnessed that the student's engagement is enhanced by face-to-face interaction with Baldi (Bosseler & Massaro, 2003; Massaro & Light, 2003).

Our software can generate a talking face (with an optional body) in real time on a personal computer, and Baldi is able to say anything at

any time in our applications. Baldi can be thought of as a puppet controlled by a set of strings that move and modify its appearance. In the algorithm for the synthesis of visible speech, each speech segment is specified with a target value for each string or what we call a facial control parameter. Critical components of the animation are to blend the successive segments together and to implement coarticulation, which is defined as changes in the articulation of a speech segment due to the influence of neighboring segments. The algorithm for animating coarticulation is based on a model of speech production using rules that describe the relative dominance of the characteristics of the speech segments (Cohen & Massaro, 1993; Massaro, 1998).

A central and somewhat unique quality of our work is the empirical evaluation of the visible speech synthesis, which has been carried out hand-in-hand with its development. The quality and intelligibility of Baldi's visible speech have been repeatedly modified and evaluated to accurately simulate naturally talking humans (Massaro, 1998). The gold standard we use is how well Baldi compares to a real person. Given that viewing a natural face improves speech perception, we determine the extent to which Baldi provides a similar improvement. We repeatedly modify the control values of Baldi in order to meet this criterion. We modify some of the control values by hand and also use data from measurements of real people talking (Cohen, Massaro, & Clark, 2002; Ouni, Massaro, Cohen, & Young, 2003). Versions of Baldi now speak a variety of languages, including Arabic (Badr; Ouni et al., 2003), Spanish (Baldero), Mandarin (Bao), Italian (Baldini; Cosi et al., 2002), German (Baltheasar), and French (Baladin).

### Value of Face-to-Face Input

There are several reasons why the use of auditory and visual information from an accurate talking head like Baldi is so successful, and why it holds so much promise for language tutoring (Massaro, 1998). These include (a) the information value of visible speech, (b) the robustness of visual speech, (c) the complementarity of auditory and visual speech, and (d) the optimal integration of these two sources of information. We review evidence for each of these properties in this chapter and begin by describing an experiment illustrating how facial information increases recognition and memory for linguistic input.

#### *Information Value of Visible Speech*

In a series of experiments, we asked 71 typical college students to report the words of sentences presented in noise (Jesse, Vrignaud, & Massaro, 2000/01). On some trials, only the auditory sentence was presented (unimodal condition). On some other trials, the auditory sentence was accompanied by Baldi, which was appropriately aligned with the auditory sentence (bimodal condition). The test items consisted of 65 meaningful sentences from the database of Bernstein and Eberhardt (1986), for example, "We will eat lunch out." The sentences were three, four, and five syllables in length and consisted of 43 statements, 17 questions, and 5 imperatives.

Figure 10.2 gives the proportion of words correctly reported for the unimodal and bimodal conditions for each of the 71 participants. As can be seen in figure 10.2, the talking face facilitated performance for everyone. Performance was more than doubled for those participants performing particularly poorer given auditory speech alone. Although a unimodal visual condition was not included in the experiment, we know that participants would have performed significantly lower than the unimodal auditory condition. Thus, the combination of auditory and visual speech has been described as synergistic because their combination can lead to accuracy that is significantly greater than accuracy on either modality alone.

Similar results are found when noise-free speech is presented to persons with limited hearing. Erber (1972) tested three populations of children (adolescents and young teenagers) strictly defined by their amount of hearing: normal hearing (NH), severely impaired (SI), and profoundly deaf (PD). The test consisted of a videotaped speaker pronouncing the eight consonants /b/, /d/, /g/, /p/, /t/, /k/, /m/, and /n/ spoken in a bisyllabic context /aCa/, where C refers to one of the eight consonants. Although all three groups benefited from seeing the face of the speaker, the SI group revealed the largest performance gain in the bimodal condition relative to either of the unimodal conditions

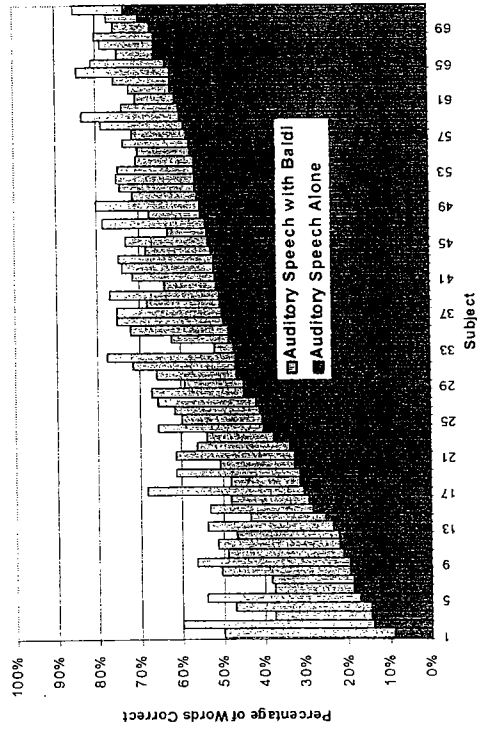


Figure 10.2. Proportion of words correctly reported for auditory speech alone and auditory speech with Baldi conditions for each of the 71 participants in the task (after Jesse et al., 2000/01).

(Massaro & Cohen, 1999). The NH group had very good auditory information, so the face could not contribute much, whereas the PD group had very poor auditory information, so the voice could not contribute much. The SI group, on the hand, had a reasonable degree of both auditory and visual information. As noted in the following discussion of complementarity and optimal integration, perception of speech can be very good when some hearing is present and the face of the speaker can be seen.

Finally, the strong influence of visible speech is not limited to situations with degraded auditory input, whether due to a noisy signal or hearing loss. Even with high-quality auditory input, a perceiver's recognition of an auditory-visual syllable can reflect the contribution of both sound and sight. For example, if the ambiguous auditory sentence "My bab pop me poo brive" is paired with the visible sentence "My gag kok me koo grive," the perceiver is likely to hear "My dad taught me to drive." Two ambiguous sources of information are combined to create a meaningful interpretation (Massaro, 1998).

#### *Robustness of Visible Speech*

Empirical findings indicate that speechreading, or the ability to obtain speech information from the face, is robust; that is, perceivers are fairly good at speechreading in a broad range of viewing conditions. To obtain information from the face, the perceiver does not have to fixate

directly on the talker's lips but can be looking at other parts of the face or even somewhat away from the face (Smeele, Massaro, Cohen, & Sittig, 1998). Furthermore, accuracy is not dramatically reduced when the facial image is blurred (because of poor vision, e.g.), when the face is viewed from above, below, or in profile, or when there is a large distance between the talker and the viewer (Massaro, 1998; Munhall & Vatikiotis-Bateson, 2004). These findings indicate that speechreading is highly functional in a variety of nonoptimal situations. The robustness of the influence of visible speech is illustrated by the fact that people naturally integrate visible speech with audible speech even when the temporal occurrence of the two sources is displaced by about a one fifth of a second (Massaro & Cohen, 1993). Light and sound travel at different speeds, and the dynamics of their corresponding sensory systems also differ (the retina transduces a visual stimulus much more slowly than the cochlea transduces an auditory one). Thus, a cross-modal integration should be relatively immune to small temporal asynchronies (Massaro, 1998, ch. 3).

#### *Complementarity of Auditory and Visual Information*

A visual talking head allows for complementarity of auditory and visual information. Auditory and visual information are complementary when one of these sources is most informative in those cases in which the other is weakest. Because of this, a speech distinction between segments is differentially supported by the two sources of information. That is, two segments that are robustly conveyed in one modality are relatively ambiguous in the other modality (Massaro & Cohen, 1999). For example, the difference between /ba/ and /da/ is easy to see but relatively difficult to hear. On the other hand, the difference between /ba/ and /pa/ is relatively easy to hear but very difficult to discriminate visually. The fact that two sources of information are complementary makes their combined use much more informative than would be the case if the two sources were noncomplementary, or redundant (Massaro, 1998, ch. 14).

#### *Optimal Integration of Auditory and Visual Speech*

The final value afforded by a visual talking head is that perceivers combine or integrate the auditory and visual sources of information in an optimally efficient manner (Massaro, 1987; Massaro & Cohen, 1999; Massaro & Stork, 1998). There are many possible ways to treat two sources of information: use only the most informative source, average the two sources together, or integrate them in such a fashion that both sources are used but that the least ambiguous source has the most influence. Perceivers in fact integrate the information available from

each modality to perform as efficiently as possible (Massaro, 1998). The best evidence for optimal integration comes from an important manipulation that systematically varies the ambiguity of each source of information in terms of how much it resembles each syllable (Massaro, 1998). In a series of experiments, the properties of the auditory stimulus were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, properties of our animated face were varied to give a continuum between visual /ba/ and /da/. Five levels of audible speech varying between /ba/ and /da/ were crossed with five levels of visible speech varying between the same alternatives. In addition, the audible and visible speech also were presented alone for a total of  $25 + 5 + 5 = 35$  independent stimulus conditions. This so-called expanded factorial design has been used with 82 participants who were repeatedly tested, giving 24 observations at each of the 35 stimulus conditions for each participant. These results have served as a database for testing models of pattern recognition (Massaro, 1998).

The proportion of /da/ responses for each of the stimulus conditions was computed for each participant. The results of one representative participant are presented in figure 10.3 to illustrate the nature of the data analysis and model testing. Figure 10.3 gives the observed (points) proportion of /da/ judgments as a function of the observed (points) stimuli in the unimodal and bimodal conditions. Although figure 10.3 might seem somewhat intimidating at first glance, a graphical analysis of this nature can dramatically facilitate understanding of the underlying processes. Only two levels of visible speech are shown in the graph for pedagogical purposes. Notice that the columns of points are spread unevenly along the x-axis. The reason is that they are placed corresponding to the influence of the auditory speech (at a value equal to the marginal probability of a /da/ judgment for each auditory level of the independent variable). This spacing thus reflects relative influence of adjacent levels of the auditory condition.

The single modality (unimodal) auditory curve (indicated by the open circles) shows that the auditory speech had a large influence on the judgments. More generally, the degree of influence of this modality when presented alone would be indicated by the steepness of the response function. The unimodal visual condition is plotted at .5 (which is considered to be completely neutral) on the auditory scale. The influence of the visual speech when presented alone is indexed by the vertical spread between the two levels of the visual condition.

The other points give performance for the auditory-visual (bimodal) conditions. This graphical analysis shows that both the auditory and the visual sources of information had a strong impact on the identification judgments. The likelihood of a /da/ identification increased as the auditory speech changed from /ba/ to /da/, and analogously for

**Evaluation of How Two Sources Are Used**

To address how the two sources of information are used, three points are circled in figure 10.3 to highlight the conditions involving the fourth level of auditory information (A4) and the third level of visual information (V3). When presented alone,  $P(/da/ | A4)$  (i.e., the probability of perceiving /da/ at A4) and  $P(/da/ | V3)$  are both about .8. When these two stimuli occur together,  $P(/da/ | A4 V3)$  is about .95. This so-called synergistic result (the bimodal is more extreme than either unimodal response proportion) does not seem to be easily explained by either the use of a single modality during a given presentation or a simple averaging of the two modalities. In order to systematically evaluate theoretical alternatives, however, formal models must be proposed and tested against all of the results, not just selected conditions. It therefore is useful to understand how one formalizes two competing models and test them against the results.

According to nonintegration models, any perceptual experience results from only a single sensory influence. Thus, the pattern recognition of any cross-modal event is determined by only one of the modalities, even though the influential modality might vary from one categorization event to the next. This idea is in the tradition of selective attention theories according to which only a single channel of information can be processed at any one time (Pashler, 1998). According to the single-channel model (SCM), only one of the two sources of information determines the response on any given trial. Formalization of the SCM is given in Massaro (1998).

*The Fuzzy Logical Model of Perception*

According to integration models, multiple sensory influences are combined before categorization and perceptual experience. The fuzzy logical model of perception (FLMP) assumes that the visible and audible speech signals are integrated. Before integration, however, each source is evaluated (independently of the other source) to determine how much that source supports various alternatives. The integration process combines these support values to determine how much their combination supports the various alternatives. The perceptual outcome for the perceiver will be a function of the relative degree of support among the competing alternatives.

Figure 10.4 illustrates three major operations during pattern recognition in the FLMP. Features are first independently evaluated (as sources of information) in terms of the degrees to which they match specific object prototypes in memory. Each feature match is represented by a common metric of fuzzy logic truth-values that range from 0 to 1 (Zadeh, 1965). In the second operation, the feature values corresponding to a given prototype are multiplied to yield an overall (absolute) goodness of match for that alternative. Finally, the goodness of

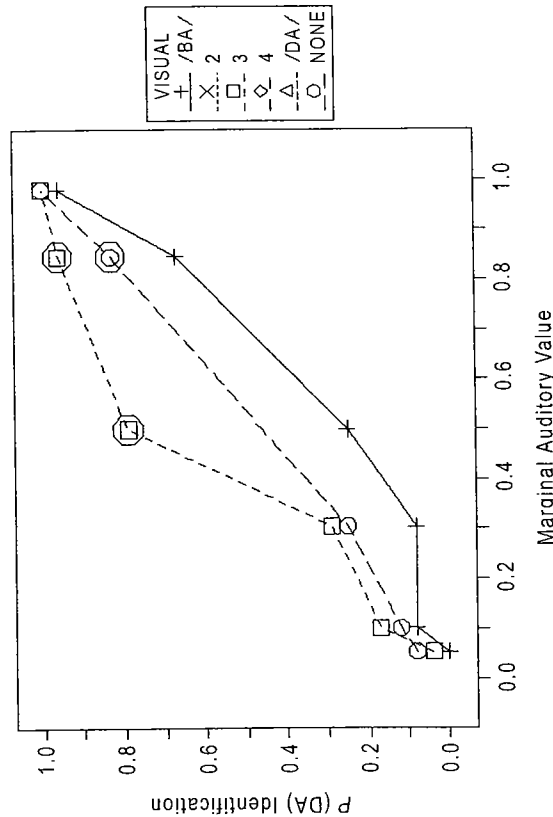


Figure 10.3. The points give the observed proportion of /da/ identifications in the unimodal and factorial auditory-visual conditions as a function of the five levels of synthetic auditory and visual speech varying between /ba/ and /da/. Only two levels of visible speech are shown in the graph for pedagogical purposes. The columns of points are placed at a value corresponding to the marginal probability of a /da/ judgment for each auditory level on the independent variable. The auditory alone conditions are given by the open circles. The unimodal visual condition is plotted at .5 (completely neutral) on the auditory scale. The predictions of the FLMP gave an RMSD of .051. (RMSD is Root Mean Square Deviations) Results shown are for participant 30.

the visible speech. The curves across changes in the auditory variable are relatively steep and also spread out from one another with changes in the visual variable. By these criteria, both sources had a large influence in the bimodal conditions.

Finally, the auditory and visual effects are not additive in the bimodal condition, as demonstrated by a significant auditory-visual interaction. The interaction is indexed by the change in the spread among the two bimodal curves across changes in the auditory variable. This vertical spread among the two curves is much greater toward the middle than at the ends of the auditory continuum. It means that the influence of one source of information is greatest when the other source is neutral or ambiguous. To understand multimodal speech perception, it is essential to understand how the two sources of information are used in perception. This question is addressed in the next section.

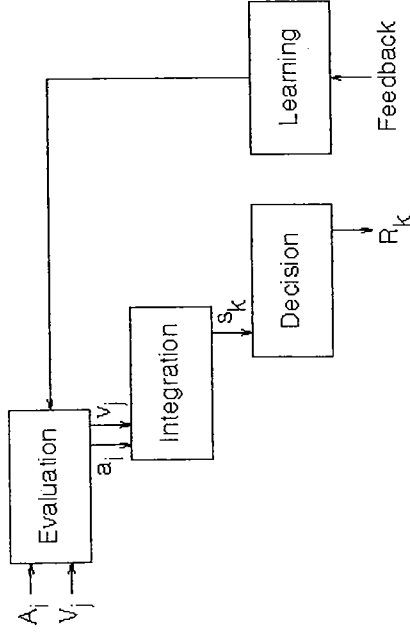


Figure 10-4. Schematic representation of the FLMP to include learning with feedback. The three perceptual processes are shown to proceed left to right in time, to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase letters. Auditory information is represented by  $A_i$  and visual information by  $V_j$ . The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters  $a_i$  and  $v_j$ ). These sources are then integrated to give an overall degree of support,  $s_k$ , for each speech alternative  $k$ . The decision operation maps the outputs of integration into some response alternative,  $R_k$ . The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The feedback during learning is assumed to tune the prototypical values of the features used by the evaluation process.

match for each alternative is compared relatively to the sum of the support for all relevant alternatives (the relative goodness rule; Massaro, 1998).

To explain pattern recognition, representations in memory are an essential component. The current stimulus input has to be compared to the pattern recognizer's memory of previous patterns. One type of memory is a set of summary descriptions of the meaningful patterns. These summary descriptions are called prototypes, and they contain a description of features of the pattern. The features of the prototype correspond to the ideal values that an exemplar should have if it is a member of that category. To recognize a speech segment, the evaluation process assesses the input information relative to the prototypes in memory.

The FLMP takes a strong stance on the question of discrete versus continuous information processing. Information input to a stage or output from a stage is continuous rather than discrete. Furthermore,

the transmission of information from one stage to the next is assumed to occur continuously rather than discretely. The three processes shown in figure 10.4 are offset to emphasize their temporal overlap. Evaluated information is passed continuously to integration while additional evaluation is taking place. Although it is logically the case that some evaluation must occur before integration can proceed, the processes are assumed to overlap in time. Similarly, integrated information is continuously made available to the decision process.

Given the FLMP framework, we are able to make an important distinction between "information" and "information processing." The sources of information from the auditory and visual channels make contact with the perceiver at the evaluation stage of processing. The reduction in uncertainty effected by each source is defined as information. In the fit of the FLMP, for example, the degree of support for each alternative from each modality corresponds to information. The predicted response probability in the unimodal condition is predicted to be a direct measure of the information given by that stimulus. These values represent how informative each source of information is. Information processing refers to how the sources of information are processed. In the FLMP, this processing is described by the evaluation, integration, and decision stages.

Across a range of studies comparing specific mathematical predictions (Massaro, 1988, 1989, 1998; Massaro, Weidon, & Kitzis, 1991), the FLMP has been more successful than other competitor models in accounting for the experimental data (Massaro, 1989, 1998; Massaro & Friedman, 1990). It is worthwhile to address two important issues related to the integration of auditory and visual speech.

### Why Integration

One might question why perceivers integrate several sources of information when just one of them might be sufficient. Most of us do reasonably well in communicating over the telephone, for example. Part of the answer might be grounded in our ontology. Integration might be so natural for adults even when information from just one sense would be sufficient because, during development, there was much less information from each sense and therefore integration was all the more critical for accurate performance (Lewkowicz & Kraebel, 2004).

### Underlying Neural Mechanism

If we really want to understand speech processing in deaf as well as hearing individuals, a natural question concerns the neural mechanism underlying the integration algorithm specified in the FLMP. An important set of observations from single-cell recordings in the cat's brain could be interpreted in terms of integration of the form specified by the FLMP (Meredith, 2004; Stein, Jiang, & Stanford, 2004; Stein & Meredith, 1993). A single hissing sound or a light spot can activate neurons in the



superior colliculus. A much more vigorous response is produced, however, when both signals are simultaneously presented from the same location. These results parallel the outcomes we have observed in unimodal and bimodal speech perception.

As proven elsewhere, the FLMP is mathematically equivalent to Bayes's theorem (Massaro, 1998, ch. 4), which is an optimal method for combining two sources of evidence to test among hypotheses. Anastasio and Patton (2004) propose that the brain can implement a computation analogous to Bayes's rule and that the response of a neuron in the superior colliculus is proportional to the posterior probability that a target is present in its receptive fields, given its sensory input. The authors also assume that the visual and auditory inputs are conditionally independent given the target, corresponding to our independence assumption at the evaluation stage. They show that the target-present posterior probability computed from the impulses from the auditory and visual neurons is higher given sensory inputs of two modalities than it is given input of only one modality, analogous to the synergistic outcome of the FLMP.

#### **A Universal Principle and Its Implications for Language Learning**

The FLMP has proven to be a universal principle of pattern recognition (Campbell, Schwarzer, & Massaro, 2001; Massaro, 1998, 2002; Massaro, Cohen, Campbell, & Rodriguez, 2001; Movellan & McClelland, 2001). In multisensory texture perception, for example, there appears to be no fixed sensory dominance by vision or haptics, and the bimodal presentation yields higher accuracy than either of the unimodal conditions (Lederman & Klatzky, 2004). In many cases, these sources of information are ambiguous and any particular source alone does not usually specify completely the appropriate interpretation. Parenthetically, it should be emphasized that these processes are not necessarily conscious or under deliberate control. Most important, we have found that typically developing children integrate information from the face and voice (Massaro, 1984, 1987, 1998) as well as do deaf and hard-of-hearing children (Massaro & Cohen, 1999) and autistic children (Massaro & Bosseler, 2003; Williams, Massaro, Peel, Bosseler, & Suddendorf, 2004).

Our early research indicated that preschool as well as school children integrate auditory and visual speech (Massaro, 1984). More recently, we have shown that both hard-of-hearing children and autistic children appear to integrate information from the face and the voice. Massaro and Bosseler (2003) tested whether autistic children integrate information in the identification of spoken syllables. An expanded factorial design was used in which information from the face and voice was presented either unimodally or bimodally, and either consistent with one another or not. After training the children in speechreading to enhance the influence of visible speech from the face, the identification task

was repeated. Children behaved similarly in the two replications, except for a larger influence of the visible speech after training in speechreading. The FLMP gave a significantly better description of performance than the SCM, supporting the interpretation that autistic children integrate vocal and facial information in speech perception.

There is evidence that hard-of-hearing and deaf children also integrate auditory and visual speech. Erber (1972) tested three populations of children under auditory, visual, and bimodal conditions. The FLMP was applied to the results of all three groups and gave an excellent description of the identification accuracy and confusion errors of all three groups of children (Massaro, 1998, ch. 14). Erber's results also reveal a strong complementarity between the audible and visible modalities in speech, which is discussed more fully in Massaro (1998, ch. 14). These results from typically developing children as well as deaf and hard-of-hearing and autistic children indicate that multisensory environments should be ideal for speech and language learning.

#### **LANGUAGE TUTORING BY COMPUTER-BASED INSTRUCTION WITH BALDI**

Computer-based instruction is an emerging method to train and develop vocabulary knowledge for both native- and second-language learners (Druin & Hendler, 2000; Wood, 2001) and individuals with special needs (Barker, 2003; Heimann, Nelson, Tjus, & Gilberg, 1995; Moore & Calvert, 2000). An incentive to employing computer-controlled applications for training is the ease with which individually-tailored instruction, automated practice and testing, feedback, and branching can be programmed. Another valuable component of computer-based instruction is the potential to present multiple sources of information, such as text, sound, and images in sequence or in parallel (Chun & Plass, 1996; Dubois & Vial, 2000). Incorporating text and visual images of the vocabulary to be learned along with the actual definitions and spoken words facilitates learning and improves memory for the target vocabulary. Dubois and Vial (2000), for example, found an increase in recall of second-language vocabulary when training consisted of combined presentations of spoken words, images, written words, and text relative to only a subset of these formats.

Computer-based instruction can easily be made available to the child most hours of a day and most days of the year. Having continual access to the instruction is valuable because learning and retention are positively correlated with the time spent learning. Take, for example, an autistic boy who has irregular sleep patterns. He could conceivably wake in the middle of the night and participate in language learning with Baldi as his friendly guide. The instruction would be tailored exactly to his strengths and needs. Other benefits of this program could



possible for the students to (1) observe the words being spoken by a realistic embodied conversational agent (Baldi), (2) experience the word as spoken as well as written, (3) see visual images of referents of the words, (4) click on or point to the referent or its spelling, (5) hear themselves say the word followed by a correct pronunciation, (6) spell the word by typing, and (7) observe and respond to the word used in context. Table 10.1 gives a description of the eight application exercises in the Language Wizard/Tutor. We now justify the importance of the direct teaching of vocabulary.

### Essential Role for Vocabulary Knowledge in Language Development

Vocabulary knowledge is central for understanding the world and for language competence both in spoken language and in reading (Gupta & MacWhinney, 1997). Empirical evidence indicates that very young normally-developing children more easily form conceptual categories when category labels are available than when they are not (Waxman, 2002). Once the child knows about 150 words, there is a sudden increase in the rate at which new words are learned and the emergence of grammatical skill (Marchman & Bates, 1994). Even children experiencing language delays because of specific language impairment benefit once this level of word knowledge is obtained. Vocabulary knowledge is positively correlated with both listening and reading comprehension (Anderson & Freebody, 1981; Stanovich, 1986; Wood, 2001) and predicts overall success in school (Vermeer, 2001). It follows that increasing the pervasiveness and effectiveness of vocabulary learning offer a promising opportunity for improving conceptual knowledge and language competence for all individuals, whether or not they are disadvantaged because of sensory limitations, learning disabilities, or social condition.

### Validity of the Direct Learning of Vocabulary

There are important reasons to justify the need for direct teaching of vocabulary. Although there is little emphasis on the acquisition of vocabulary in typical school curricula, research demonstrates that some direct teaching of vocabulary is essential for appropriate language development in normally-developing children (Beck et al., 2002). Contrary to a common belief that learning vocabulary is a necessary outcome of reading in which new words are experienced in a meaningful context, context seldom disambiguates the meaning of a word completely. As an example, consider a passage from *The Fir Tree* by Hans Christian Andersen:

Then two servants came in rich livery and carried the Fir Tree into a large and splendid drawing-room. Portraits were hanging on the walls, and near the white porcelain stove stood two large Chinese vases with lions on the covers.

Table 10-1: Description of the Eight Application Exercises in the Language Wizard/Player

Application Exercise	Description
Pretest	Baldi instructs the student to "click on the zucchini," and the student is required to drag the computer mouse over the item that was just presented and click on it. Feedback can be given about the student's response via a happy or sad face. Items can be randomly presented a variable number of times.
Presentation	One image becomes highlighted, and Baldi tells the student, "This is a zucchini" (for example). The written label of the vocabulary item can appear on the screen below the canvas of images. Baldi then instructs the student, "Show me the zucchini," and the student is required to drag the computer mouse over the highlighted image and click on it. Feedback can be given. This is to reinforce that the student knew which image was being described. Items can be randomly presented a variable number of times.
Identification	Baldi instructs the student, "Click on the zucchini" (for example), and the student is required to drag the computer mouse over the item that was just presented and click on it. Feedback can be given. If the student chose the wrong item, the correct item is highlighted and Baldi tells the student that the word they chose was not the zucchini and that the item that is highlighted is the zucchini. Items can be randomly presented a variable number of times.
Reading	The written text of all of the vocabulary items is presented below the images. Baldi instructs the student to click on the word corresponding to the highlighted image. Feedback can be given. Items can be randomly presented a variable number of times.
Spelling	One of the images is highlighted while Baldi asks the student to type the corresponding word. Feedback can be given. If the student is incorrect, the correct spelling of the vocabulary item appears above the student's attempt and Baldi reads the word and spells it out to the student.
Imitation	One of the images is highlighted, and Baldi names the item. The student is instructed to repeat what Baldi had just said after the tone. The student says the word, and his or her voice is recorded and played back. Baldi can then say the word again to reinforce the child's pronunciation. Items can be randomly presented a variable number of times.

(continued)

a student can profit from the repeated experience of practicing new words in multiple contexts during the direct teaching of vocabulary.

The Language Wizard/Tutor with Baldi encompasses and instantiates the developments in the pedagogy of how language is acquired, remembered, and used. Direct teaching of vocabulary by computer software is possible, and an interactive multimedia environment is ideally suited for this learning (Wood, 2001). The Language Tutor provides a learning platform that allows optimal conditions for learning and the engagement of fundamental psychological processes such as working memory, the phonological loop, and the visual-spatial scratchpad (Atkins & Baddeley, 1998). Evidence by Baddeley and colleagues (Baddeley, Gathercole, & Papagno, 1998; Evans et al., 2000) supports a strategy of centering vocabulary learning in spoken language dialogs. There is also some evidence that reading aloud activates brain regions that are not activated by reading silently (Beminger & Richards, 2002). Thus, the imitation and elicitation activities in the Language Tutor should reinforce learning of vocabulary and grammar. Experimental tests of the effectiveness of the Language Wizard/Tutor with hard-of-hearing children are described next.

**Effectiveness of Language Wizard/Tutor**

Hard-of-hearing children have significant delays in both spoken and written vocabulary knowledge (Breslaw et al., 1981; Holt et al., 1997). One reason is that these children tend not to overhear other conversations because of their limited hearing and are thus shut off from an opportunity to learn vocabulary. These children often do not have names for specific things and concepts and therefore communicate with phrases such as "the window in the front of the car," "the big shelf where the sink is," or "the step by the street" rather than "windshield," "counter," or "curb" (Barker, 2003). In an initial independent evaluation carried out by Barker (2003), 13 teachers successfully used the Language Wizard to compose individually tailored lessons for their hard-of-hearing students, and the students learned and retained new vocabulary from these lessons.

Students photographed surroundings at home. Pictures of 10-15 objects were then incorporated in the lessons. Students practiced the lessons about 10 minutes a day until they reached 100% on the posttest. They then moved on to another lesson. They were also retested about 1 month after each successful (100%) posttest. Ten girls and nine boys 8-14 years old participated; 16 were hard-of-hearing children, and 3 were hearing children.

Figure 10.6 gives the average results of these lessons for three stages: pretest, posttest, and retention after 30 days. The items were classified as known, not known, and learned. Known items are those children knew on the pretest before the first lesson. Not known items are those

**Table 10-1: (continued)**

Application Exercise	Description
Elicitation	One of the images is highlighted, and Baldi asks the student to name it. Independent of the student's production response, Baldi can then say the word again to reinforce the child's pronunciation. Items can be randomly presented a variable number of times.
Posttest	Baldi instructs the student, "Click on the zucchini," and the student is required to drag the computer mouse over the item that was just presented and click on it. Feedback can be given about the student's response via a happy or sad face. Items can be randomly presented a variable number of times.

Most of the words in this passage are not disambiguated by context. The meaning of livery, portraits, porcelain, and vases, for example, cannot be determined from the context of the story alone. Research by Beck et al. (2002) and Baker, Simmons, and Kameenui (1995) provides some evidence that hearing children more easily acquire new vocabulary by direct intentional instruction than by other incidental means (see also McKeown, Beck, Omanson, & Pople, 1985; Pany & Jenkins, 1978; Stahl, 1983). Although there does not appear to be any analogous research with deaf and hard-of-hearing children, the same advantage of direct instruction presumably would exist for them or other children with language challenges.

Direct instruction is also valuable because knowing a word is not an all-or-none proposition. A single experience with a word (even if the correct meaning of the word is comprehended) is seldom sufficient for mastering that word. Acquiring semantic representations appears to be a gradual process that can extend across several years (McGregor, Friedman, Reilly, & Newman, 2002). Words are complex multidimensional stimuli, and a person's knowledge of the word will not be as complete or as accurate as its dictionary entry. Semantic naming errors are more likely to occur with those items that have less embellished representations. Thus, it is important to overtrain or continue vocabulary training after the word is apparently known and to present the items in a variety of contexts in order to develop rich representations. Picture naming and picture drawing are techniques that can be used to probe and reinforce these representations (McGregor et al., 2002). Qian (2002) found that the dimension of vocabulary depth (as measured by synonymy, polysemy, and collocation) is as important as that of vocabulary size in predicting performance on academic reading. Therefore,

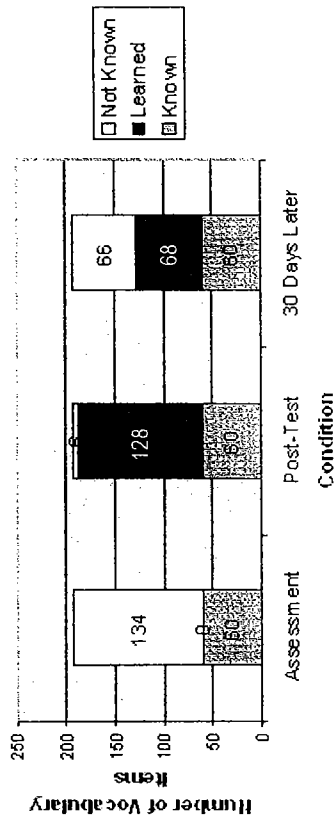


Figure 10-6. Results of word learning at the Tucker-Maxon Oral School using the Language Wizard/Tutor. The results showed significant vocabulary learning, with more than 50% retention of new words after 30 days.

children were unable to identify in the pretest. Learned items are those not-known items identified correctly in the posttest. Students knew about half of the items without any learning; they learned the other half of the items and retained about one half of the newly learned items when retested 30 days later.

Since no control groups were used in the evaluation described above, it was possible the children were learning the words outside of the Language Tutor environment. In addition, the results do not give the rate of vocabulary acquisition. It would also be valuable to measure production of the words, given that only identification was measured previously.

To study these questions, we used a within-student multiple baseline design (Baer, Wolf, & Risley, 1968; Horner & Baer, 1978), where certain words were continuously tested while others were tested and trained (Massaro & Light, 2004a). Although the teachers and speech therapists agreed not to use these words during our investigation, it is still possible that the words could be learned outside of the Language Tutor environment. The single-student multiple-baseline design monitors this possibility by providing a continuous measure of the knowledge of words that are not being trained. Thus, any significant differences in performance on the trained words and untrained words can be attributed to the Language Tutor training program itself rather than some other factor. In addition, this design tracks the rate of learning of each child.

Eight hard-of-hearing children, two males 6 and 7 years old and six females 9 and 10 years old, were recruited with parental consent from the Jackson Hearing Center in Palo Alto, California. Table 10.2 gives a description of the children and their hearing. Using the Language Wizard, the experimenter developed lessons with a collection of

Table 10-2. Age of the Participants at the Midpoint of the Study and Individual and Average Aided Auditory Device Thresholds (dB HL) at Four Frequencies for the Eight Students Studied in Massaro and Light (2004a)

S#	Age (year; month)	500 Hz	1,000 Hz	2,000 Hz	4,000 Hz	PTA	ULE	URE
1	7; 2	40	35	47	55	41	78	80
2	6; 11	35	30	35	43	33	80	85
3	10; 7	25	35	40	45	33	35	35
4	9; 3	30	33	45	68	36	95	42
5	11; 0	40	35	40	50	38	—	—
6	10; 0	50	52	55	60	52	95	95
7	9; 4	25	15	25	35	21	90	80
8	9; 11	30	35	40	60	35	110	60
Mean	9; 3	34	34	41	52	36	83	68

Participants 1 and 2 were in grade 1, and the others were in grade 4. Participant 7 had a cochlear implant, and the seven other children had bilateral hearing aids, except for participant 8, who had just one aid. The participant numbers (S#) correspond to those in the results; PTA is pure tone average; ULE and URE are unaided thresholds for left and right ears, respectively (which are not available for S5).

vocabulary items customized to each student. Each collection consisted of 24 items to provide three lessons of eight items for each child. Images of the vocabulary items were shown on the screen next to Baldi as he spoke. As can be seen in table 10.1, one exercise asked the child to respond to Baldi's instructions such as "click on the cabbage" by clicking on the item. Other exercises asked the child to recognize or type the written word. The production exercises asked the child to repeat the word after Baldi named the highlighted image or to name the image prior to Baldi naming it.

Figure 10.7 gives the accuracy of identification and production for one of the eight students. These results are typical because the outcome was highly consistent across the eight students. Identification performance was better than production because a child would be expected to recognize an object in order to pronounce it correctly. There was little knowledge of the test items without training, even though these items were repeatedly tested for many days. Once training began on a set of items, performance improved quickly until asymptotic knowledge was obtained. This knowledge did not degrade after training on these words ended and training on other words took place.

A reassessment test given about 4 weeks after completion of the experiment revealed that the students retained the items that were learned. The results show that the Language Tutor application is effective in teaching new vocabulary, there is a fast rate of learning given

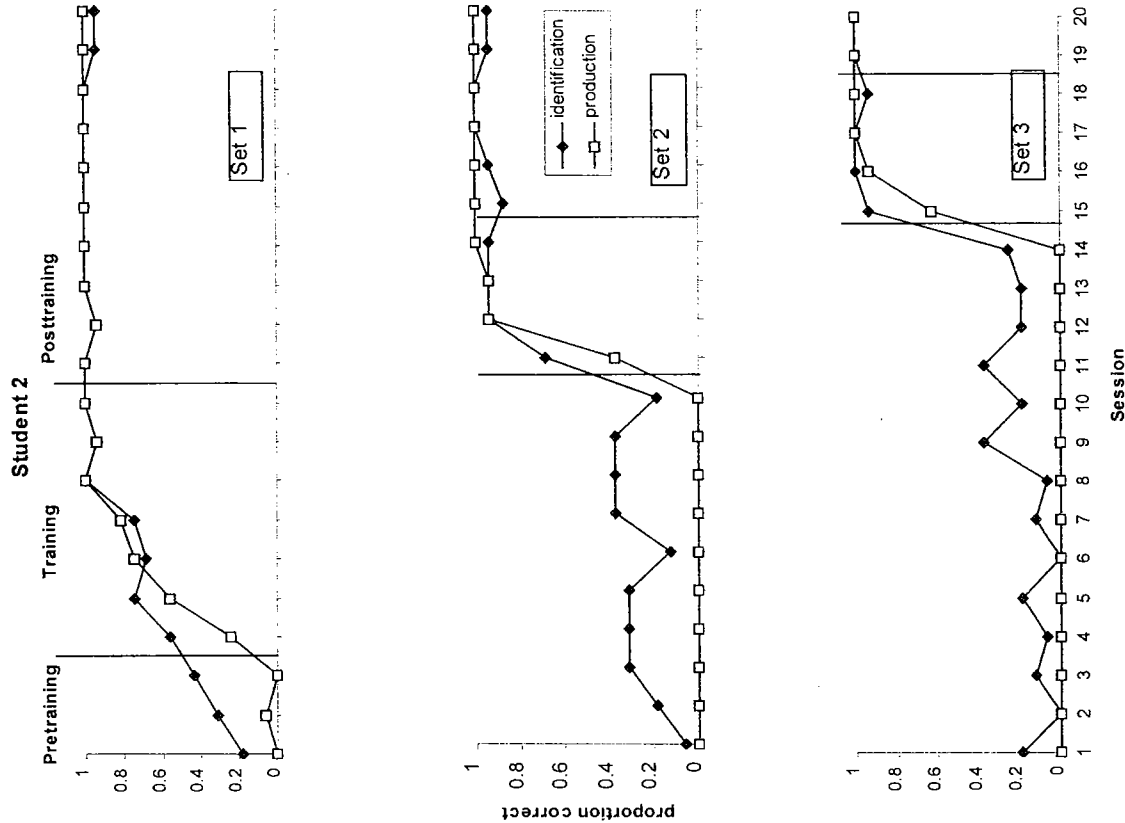


Figure 10-7. Proportion of correctly identified items (diamonds) and correctly produced items (squares) across the testing sessions for student 2. The training occurred between the two vertical bars. The figure illustrates that once training was implemented identification performance increased dramatically, and remained accurate without further training.

instruction, and the knowledge is retained at least a month after training is completed.

**Development and Evaluation of a Speech Training Tutor**

We have extended previous approaches to spoken language intervention for deaf children (e.g., Ling, 1976; Ling & Ling, 1978) by viewing speech learning as a multisensory experience (Massaro et al., 1999, Massaro, Cohen, Tabain, Beskow, & Clark, in press), and we have tested the idea that Baldi can function effectively as a language tutor to teach speech perception and production. Baldi has a tongue, hard palate, and three-dimensional teeth, and his internal articulatory movements have been trained with electropalatography and ultrasound data from natural speech (Cohen et al., 1998; Massaro et al., in press). Although previous approaches have used palatometry (Fletcher, Dagenais, & Critz-Crosby, 1991) and electropalatography (Hardcastle & Gibbon, 1997) as a form of visual feedback, Baldi can display a more representative view of the actual articulation. Baldi can demonstrate articulation by illustrating a midsagittal view, or the skin on the face can be made transparent to reveal the internal articulators, as shown in figure 10.8. In addition to simply showing the actual articulation, the area of contact between the tongue and palate and teeth can be highlighted.

The orientation of the face can be changed to display different viewpoints while speaking, such as a side view or a view from the back of the head (Massaro, 1998). As an example, a unique view of Baldi's internal articulators can be presented by rotating the exposed head and vocal tract to be oriented away from the student. It is possible that this back-of-head view would be much more conducive to learning language production. The tongue in this view moves away from and toward the student in the same way as the student's own tongue would move. This correspondence between views of the target and the student's articulators might facilitate speech production learning. One analogy is the way one might use a map. We often orient the map in the direction we are headed to make it easier to follow (e.g., turning right on the map is equivalent to turning right in reality).



Figure 10-8. Various views of Baldi that can be used in speech and language tutoring.

Baldi's auditory and visual speech can also be independently controlled and manipulated, permitting customized enhancements of the informative characteristics of speech. These features offer novel approaches in language training, permitting Baldi to pedagogically illustrate appropriate articulations that are usually hidden by the face. Baldi can be made even more informative by embellishing of the visible speech with added features. Distinguishing phonemes that have similar visible articulations, such as the difference between voiced and voiceless segments, can be indicated by vibrating the neck. Nasal sounds can be marked by making the nasal opening red, and turbulent airflow can be characterized by lines emanating from the mouth during articulation. These embellished speech cues could make the face more informative than it normally is. Based on reading research, we expected that these additional visible cues would heighten the child's awareness of the articulation of these segments and assist in the training process.

Children with hearing loss require guided instruction in speech perception and production. Some of the distinctions in spoken language cannot be heard with degraded hearing—even when the hearing loss has been compensated by hearing aids or cochlear implants. To test whether the Baldi technology has the potential to help individuals with hearing loss, Massaro and Light (2004b) carried out a speech training study. Seven students (two male and five female) from the Jackson Hearing Center and JLS Middle School in Los Altos, California, participated in the study. Table 10.3 gives a description of the children and their hearing.

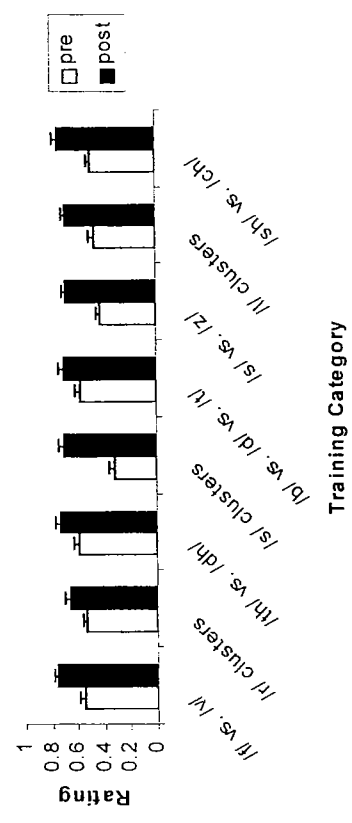
The students were trained to discriminate minimal pairs of words bimodally (auditorily and visually) and were also trained to produce

**Table 10.3: Individual and Average Unaided and Aided Auditory Thresholds (dB HL) for Four Frequencies for the Seven Students Studied in Massaro and Light (2004b)**

Student No.	Aids	Unaided Aided Auditory Thresholds (dB HL)				Age (years)
		500 Hz	1000 Hz	2000 Hz	4000 Hz	
1	Binaural	60/25	60/15	75/30	85/45	12
2	Binaural	60/28	80/25	85/30	80/30	11
3	Binaural	85/50	90/55	85/55	80/55	8
4	Binaural	45/15	60/35	65/45	60/50	11
5	CI-1- left ear	95/40	110/25	115/25	105/35	13
6	Binaural	65/30	95/30	100/35	110/70	13
7	Right ear	65/30	55/35	65/40	70/60	8
Average		68/31	79/31	84/37	84/49	11

CI, Cochlear implant.

**Production Ratings as a function of Category Involved**



**Figure 10.9. Intelligibility ratings of the pretest and posttest word productions (and standard error bars) for each of the eight training categories.**

various speech segments by visual information about how the inside oral articulators work during speech production. The articulators were displayed from different vantage points so that the subtleties of articulation could be optimally visualized. The speech was also slowed down significantly to emphasize and elongate the target phonemes, allowing for clearer understanding of how the target segment is produced in isolation or with other segments. Each student completed eight training lessons of approximately 45 minutes for a total of 6 hours of training.

Figure 10.9 shows that the students' ability to accurately perceive and produce words involving the trained segments improved from pretest to posttest. Intelligibility ratings of the posttest productions were significantly higher than pretest productions, indicating significant learning. It is always possible that some of this learning occurred independently of our program or was simply based on routine practice. To test this possibility, we assessed the students' productions 6 weeks after training was completed. Although these productions were still rated as more intelligible than the pretest productions, they were significantly lower than posttest ratings, indicating some decrement due to lack of continued participation in the training program. This is evidence that at least some of the improvement must have been due to the program.

**SUMMARY AND CONCLUSIONS**

Perceivers expertly use multiple sources of information to identify and interpret the language input. Auditory and visual speech is seamlessly

- Barker, L. J. (2003). Computer-assisted vocabulary acquisition: The CSLU vocabulary tutor in oral-deaf education. *Journal of Deaf Studies and Deaf Education, 8*, 187-198.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford Press.
- Berninger, V. W., & Richards, T. L. (2002). *Brain literacy for educators and psychologists*. San Diego: Academic Press.
- Bernstein, L. E., Demorest, M. E., Coulter, D. C., and O'Connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America, 90*, 2971-2984.
- Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception and Psychophysics, 62*, 233-252.
- Bernstein, L. E., & Eberhardt, S. P. (1986). *Johns Hopkins lipreading corpus videodisk set*. Baltimore, MD: Johns Hopkins University.
- Bosseler, A., & Massaro, D. W. (2003). Development and evaluation of a computer-animated tutor for vocabulary and language learning for children with autism. *Journal of Autism and Developmental Disorders, 33*, 653-672.
- Breslaw, P. I., Griffiths, A. J., Wood, D. J., & Howarth, C. I. (1981). The referential communication skills of deaf children from different educational environments. *Journal of Child Psychology, 22*, 269-282.
- Campbell, C. S., Schwarzer, G., & Massaro, D. W. (2001). Face perception: An information processing perspective. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 285-345). Mahwah, NJ: Lawrence Erlbaum.
- Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *Modern Language Journal, 80*, 183-198.
- Cohen, M. M., & D. W. Massaro (1993). Modeling coarticulation in synthetic visual speech. Models and Techniques in Computer Animation. In D. Thalmann and N. Magnenat-Thalmann (Eds) *Models and techniques in computer animation* (pp. 141-155). Tokyo: Springer-Verlag.
- Cohen, M. M., Beskow, J., & Massaro, D. W. (1998, December). *Recent developments in facial animation: An inside view*. Paper presented at Auditory Visual Speech Processing '98, Sydney, Australia.
- Cohen, M. M., Massaro, D. W., & Clark, R. (2002, October). Training a talking head. In *Proceedings of ICMIT'02, IEEE Fourth International Conference on Multimodal Interfaces* (pp. 499-504). Piscataway, NJ: IEEE Computer Society.
- Cosi, P., Cohen, M. M., & Massaro, D. W. (2002). Baldini: Baldi speaks Italian. In *Proceedings of 7th International Conference on Spoken Language Processing (ICSLP'02)* (pp. 2349-2352). Denver, CO.
- Denes, P. B., & Pinson, E. N. (1963). *The speech chain. The physics and biology of spoken language*. New York: Bell Telephone Laboratories.
- Druin, A., & Hendler, J. (Eds.). (2000). *Robots for kids: Exploring new technologies for learning*. San Francisco: Morgan Kaufmann.
- Dubois, M., & Vial, I. (2000). Multimedia design: The effects of relating multimodal information. *Journal of Computer Assisted Learning, 16*, 157-165.
- Erber, N. P. (1972). Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Research, 15*, 423-442.

evaluated and integrated to facilitate understanding in face-to-face communication. This behavior is accurately described by a fuzzy logical model of perception (FLMP). Given the value of face-to-face interaction, our persistent goal has been to develop, evaluate, and apply animated agents to produce realistic and accurate speech. Baldi is an accurate three-dimensional animated talking head appropriately aligned with either synthesized or natural speech. Baldi has a tongue and palate, which can be displayed by making his skin transparent. Based on this research and technology, we have implemented computer-assisted speech and language tutors for hard-of-hearing and autistic children. Our language-training program utilizes Baldi as the conversational agent, who guides students through a variety of exercises designed to teach vocabulary and grammar, to improve speech articulation, and to develop linguistic and phonological awareness. Some of the advantages of the Baldi pedagogy and technology include the popularity and effectiveness of computers and embodied conversational agents, the perpetual availability of the program, and individualized instruction. The science and technology of Baldi hold great promise in language learning, dialog, human-machine interaction, and education.

#### ACKNOWLEDGMENTS

This work was supported in part by grants from the National Science Foundation (CHALLENGE grant CDA-9726363 and grant BCS-9905176), a grant from the Public Health Service (PHS R01 DC00236), cooperative grants from the Intel Corporation and the University of California Digital Media Program (D97-04), and grants from the University of California, Santa Cruz.

#### REFERENCES

- Anastasio, T. J., & Patton, P. E. (2004). Analysis and modeling of multisensory enhancement in the deep superior colliculus. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (265-283). Cambridge, MA: MIT Press.
- Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research perspectives* (pp. 71-117). Newark, DE: International Reading Association.
- Atkins, P. W. B., & Baddeley, A. D. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics, 19*, 537-552.
- Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 1, 158-173.
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91-97.
- Baker, S. K., Simmons, D. C., & Kameenui, E. J. (1995). *Vocabulary acquisition: Synthesis of the research*. Eugene, OR: National Center to Improve the Tools of Educators.

- Evans, J. J., Wilson, B. A., Schuri, U., Baddeley, A. D., Canavan, A., Laaksonen, R., et al. (2000). A comparison of "errorless" and "trial and error" learning methods for teaching individuals with acquired memory deficits. *Journal of the International Neuropsychological Society*, 10, 67–101.
- Fletcher, S. G., Dagenais, P. A., & Critz-Crosby, P. (1991). Teaching consonants to profoundly hearing-impaired speakers using palatometry. *Journal of Speech and Hearing Research*, 34, 929–942.
- Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: St. Martin's.
- Gupta, P., & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computation and neural bases. *Brain and Language*, 59, 267–333.
- Hardcastle, W. J., & Gibbon, F. (1997). Electropalatography and its clinical applications. In M. J. Ball & C. Code (Eds.), *Instrumental clinical phonetics* (pp. 149–193). London: Whurr.
- Heimann, M., Nelson, K., Tjus, T., & Gilberg, C. (1995). Increasing reading and communication skills in children with autism through an interactive multimedia computer program. *Journal of Autism and Developmental Disorders*, 25, 459–480.
- Holt, J. A., Traxler, C. B., & Allen, T. E. (1997). *Interpreting the scores: A user's guide to the 9th Edition Stanford Achievement Test for educators of deaf and hard-of-hearing students*. Washington, DC: Gallaudet Research Institute.
- Horner, R. D., & Baer, D. M. (1978). Multiple-probe technique: A variation of the multiple baseline. *Journal of Applied Behavior Analysis*, 11, 189–196.
- Jesse, A., Vrignaud, N., & Massaro, D. W. (2000/01). The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, 5, 95–115.
- Kisor, H. (1990). *What's that pig outdoors? A memoir of deafness*. New York: Hill and Wang.
- Lederman, S. J., & Klatzky, R. L. (2004). Multisensory texture perception. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 107–122). Cambridge, MA: MIT Press.
- Lewkowicz, D. J., & Kraebel, K. S. (2004). The value of multisensory redundancy in the development of intersensory perception. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 655–678). Cambridge, MA: MIT Press.
- Ling, D. (1976). *Speech and the hearing-impaired child: Theory and practice*. Washington, DC: Alexander Graham Bell Association for the Deaf.
- Ling, D., & Ling, A. (1978). *Aural habilitation: The foundation of verbal learning in hearing-impaired children*. Washington, DC: Alexander Graham Bell Association for the Deaf, pp. 129–130, 211.
- Marchman, V., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, 21, 339–366.
- Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development*, 55, 1777–1788.
- Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.
- Massaro, D. W. (1988). Ambiguity in perception and experimentation. *Journal of Experimental Psychology: General*, 117, 417–421.

- Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, 21, 398–421.
- Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
- Massaro, D. W. (2000, August). From "speech is special" to talking heads in language learning. In *Proceedings of Integrating Speech Technology in the Language Learning and Assistive Interface* (pp. 153–161).
- Massaro, D. W. (2002). Multimodal speech perception: A paradigm for speech science. In B. Granstrom, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp. 45–71). Dordrecht: Kluwer.
- Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 153–176). Cambridge, MA: MIT Press.
- Massaro, D. W., Beskow, J., Cohen M. M., Fry, C. L., & Rodriguez, T. (1999). Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks. In *Proceedings of Auditory-Visual Speech Processing (AVSP'99)* (pp. 133–38). Santa Cruz, CA.
- Massaro, D. W., & Bosselet, A. (2003). Perceiving speech by ear and eye: Multimodal integration by children with autism. *Journal of Developmental and Learning Disorders*, 7, 111–144.
- Massaro, D. W., & Cohen, M. M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, 13, 127–134.
- Massaro, D. W., & Cohen, M. M. (1999). Speech perception in hearing-impaired perceivers: Synergy of multiple modalities. *Journal of Speech, Language, and Hearing Science*, 42, 21–41.
- Massaro, D. W., Cohen, M. M., Beskow, J., & Cole, R. A. (2000). Developing and evaluating conversational agents. In J. Cassell, J. Sullivan, S. Prevost, & E. Churchill (Eds.), *Embodied conversational agents* (pp. 286–318). Cambridge, MA: MIT Press.
- Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin and Review*, 8, 1–17.
- Massaro, D. W., Cohen, M. M., Tabain, M., Beskow, J., & Clark, R. (in press). Animated speech: Research progress and applications. In E. Vatiokis-Bateson, G. Bailly, & P. Perrier (Eds.), *Audiovisual speech processing*. Cambridge, MA: MIT Press.
- Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, 97(2), 225–252.
- Massaro, D. W., & Light, J. (2003, August). *Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/*. *Eurospeech 2003-Switzerland (Interspeech)*. In Proceedings of the 8th European Conference on Speech Communication and Technology, (Eurospeech 2003/Interspeech 2003) (CD-Rom, 4 pages). Geneva, Switzerland.
- Massaro, D. W., & Light, J. (2004a). Improving the vocabulary of children with hearing loss. *Volta Review*, 104(3), 141–173.
- Massaro, D. W., & Light, J. (2004b). Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research*, 47(2), 304–320.

- Massaro, D. W., & Stork, D. G. (1998). Sensory integration and speechreading by humans and machines. *American Scientist*, *86*, 236–244.
- Massaro, D. W., Weldon, M. S., & Kitzis, S. N. (1991). Integration of orthographic and semantic information in memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 277–287.
- McGregor, K. K., Friedman, R. M., Reilly, R. M., & Newman, R. M. (2002). Semantic representation and naming in young children. *Journal of Speech, Language, and Hearing Research*, *45*, 332–346.
- McKeown, M., Beck, I., Omanson, R., & Pople, M. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly*, *20*, 522–535.
- Meredith, M. A. (2004). The neural mechanisms underlying the integration of cross-modal cues: Single neurons, event-related potentials and models. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 343–355). Cambridge, MA: MIT Press.
- Moore, M., & Calvert, S. (2000). Vocabulary acquisition for children with autism: Teacher or computer instruction. *Journal of Autism and Developmental Disorders*, *30*, 359–362.
- Munhall, K., & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 177–188). Cambridge, MA: MIT Press.
- Ouni, S., Massaro, D. W., Cohen, M. M., Young, K., & Jeeze, A. (2003). Internationalization of a Talking Head. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03)*. Universitat Autònoma de Barcelona, Barcelona, Spain.
- Pany, D., & Jenkins, J. R. (1978). Learning word meanings: A comparison of instructional procedures and effects on measures of reading comprehension with learning disabled students. *Learning Disability Quarterly*, *1*, 21–32.
- Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*, 513–536.
- Smeele, P. M. T., Massaro, D. W., Cohen, M. M., & Sittig, A. C. (1998). Laterality in visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1232–1242.
- Stahl, S. (1983). Differential word knowledge and reading comprehension. *Journal of Reading Behavior*, *15*(4), 33–50.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360–406.
- Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.
- Stein, B. E., Jiang, W., & Stanford, T. R. (2004). Multisensory integration in single neurons of the midbrain. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 243–264). Cambridge, MA: MIT Press.
- Summerfield, Q. (1987). Some preliminaries to a comprehensive account of AV speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 3–51). Hillsdale, NJ: Lawrence Erlbaum.
- A Computer-Animated Tutor for Language Learning
- Tager-Flusberg, H. (2000). Language development in children with autism. In L. Menn & N. Bernstein Ratner (Eds.), *Methods for studying language production* (pp. 313–332). Mahwah, NJ: Erlbaum.
- Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*, 217–234.
- Waldstein, R. S., & Boothroyd, A. (1995). Speechreading supplemented by single-channel and multi-channel tactile displays of voice fundamental frequency. *Journal of Speech and Hearing Research*, *38*, 690–705.
- Waxman, S. R. (2002). Early word-learning and conceptual development: Everything had a name, and each name gave birth to a new thought. In U. Goswami (Ed.), *Handbook of childhood cognitive development* (pp. 102–126). Malden, MA: Blackwell Publishing.
- Williams, J. H. G., Massaro, D. W., Peel, N. J., Bosseler, A., & Suddendorf, T. (2004). *Visual-auditory integration during speech initiation in autism. Research in developmental disabilities* *25*, 559–575.
- Wood, J. (2001). Can software support children's vocabulary development? *Language Learning and Technology*, *5*, 166–201.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.